K-anonymity model for privacy-preserving soccer fitness data publishing

Rong Li[1], *Shushan* An[2,*], *Dong* Li[1], *Jian* Dong[3], *Wanjian* Bai[1], *Hongmei* Li[1], *Zhiming* Zhang[4], and *Qingyang* Lin[3]

[1]State Grid Corporation of China, Shandong, China
[2]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[3]Shandong Luneng Sports & Culture Company, State Grid Shandong Company, Shandong, China
[4]Shandong Luneng Software Technology, Shandong, China

Abstract. With the development of data mining technology, more and more researchers use the soccer fitness data to analyse the ranking of soccer athletes' and professional training. However, the direct release of soccer fitness data may leak the personal privacy of soccer athletes, so how to ensure the utility of soccer fitness data and the privacy of soccer player has become an issue. In this paper, we point out the linking attack existing in soccer fitness data, which the attackers can use the auxiliary demographic data as background information to attack the published physical data. So the attackers will map the privacy data and the athlete together. At the same time, we apply the partitioning-based and k-means clustering-based two k-anonymity algorithms to the soccer fitness data publishing to trade-offs the data utility and the personal privacy. Experimental results showed that the performance of methods is convincing.

1 Introduction

In recent years, with the rapid development of big data and Internet era, massive amounts of data are collected for various reasons by many organizations with the hope that data mining technology will extract useful knowledge from the collected data and turn it into something beneficial for the organization. There are also some obstacles on the way of the data mining. Part of the reason for that are the privacy issues of individuals. The data mining community focused on developing techniques that would enable data utility while preserving the privacy of individuals and started a popular branch of research named "Privacy Preserving Data Publishing" [1]. Rossi A *et. al.* [2] used the collected data by GPS and proposed a multidimensional approach to injury prediction in professional soccer which is based on machine learning. However, they don't consider publishing the collected fitness data to other parties to analyze more useful information.

Some recent studies [3], [4] show that, the simple technique of protecting traditional data by removing their identifiers (e.g., Name and Social Security Number) before publishing the table does not always guarantee privacy. The linking attack also exists in the

* Corresponding author: anshshan@163.com

soccer fitness data publishing. For example, a soccer player might be re-identified by joining the published data with another soccer website dataset on Height and Weight. Fig. **1** shows such an attack, where Bob's sensitive fitness data will be determined by joining the published soccer fitness data with a public soccer website data.

Soccer Website Data

Name	Height	Weight
Bob	176	64.1
Michael	169	63.8
Dave	180	77.1
John	188	81.3
Gavin	170	59.3

Soccer Fitness Data

Anonym	Height	Weight	Sensitive Fitness Data			
			Respiratory entropy	V4	HRV4	...
W	176	64.1	1.03	13.86	175.4	....
AE	176	63.8	1.1	12.48	165.13	....
Y	183	77.1	1.46	13.1	181.28	....
M	188	81.3	1.28	13.75	163.88	....
AR	170	59.3	1.12	11.25	168.2	....
AG	177	68.6	1.06	11.2	161.75	....

Fig. 1. Tables vulnerable to linking attack.

K-anonymity has been proposed to reduce the risk of this type of attack [4]. The main purpose of the k-anonymization, which has at least k-1 same tuples of each tuple, is to protect the privacy of the individual to whom the data belongs. The main contributions of this paper are summarized as follows:

- We point out the linking attack existing in soccer fitness data. The attackers can use the public soccer website data to determine a soccer player and his sensitive fitness data.
- Towards to this type of attack, this paper applies the two k-anonymity methods, Mondrian K-Anonymity and K-means Anonymity. Meantime, we conduct a comparative study of aforementioned approaches on the soccer fitness data and analyse the results of these methods. Experimental results show that these methods can preserve the soccer players' privacy and guarantee the utility of these data.

The remainder of this paper is organized as follows. Section 2 presents a brief overview of the literature on the k-anonymization. In Section 3, we formally define the k-anonymity model for traditional anonymization. Section 4 focuses on practical solutions of k-anonymity for privacy preserving data publishing. We present the experimental results in Section 5 and conclude this paper in Section 6.

2 Related work

The issue of information disclosure has been studied extensively in the framework of statistical databases. Lots of information disclosure limitation techniques have been designed for data publishing, including Sampling, Cell Suppression, Rounding, Data Swapping and Perturbation. However, these methods compromised data integrity of the tables. Sweeney [3] first introduced the k-anonymity protection model, explored related attacks and provided ways in which the attacks can be thwarted.

Numerous algorithms [5], [6], [7] have been proposed in the literature for guaranteeing k-anonymity. LeFevre $et.\ al.$ [6] introduced a class of algorithms for producing k-anonymous full-domain generalizations using two key ideas of bottom-up aggregation (rollup) along generalization dimensions and a priori computation. The work in [7] extended the above study by using a simple greedy approximation algorithm to complete the multidimensional k-anonymity.

3 Problem definition

In this section, a general model, k-anonymity, is used to define the anonymization problem for soccer fitness data. Furthermore, we consider two efficient metrics to quantify the information loss incurred by the table perturbation.

3.1 K-anonymity: A general model

Suppose a data holder wants to publish a soccer fitness data table $T(ID, D_1, \ldots, D_m, Sens)$ to some recipient for data analysis. ID is an explicit identifier, such as SSN, and it must be removed before publication. Each D_i is either a categorical or a numerical attribute. $Sens$ is a sensitive attribute.

$Definition\ 1 : (Quasi\text{-}identifier\ Attribute\ Sets)$ A $quasi\text{-}identifier(QID)$ is a minimal set of attributes $D_1, \ldots, D_m$ in table T that can be joined with external information to re-identify individual records (with sufficiently high probability), where $QID \subseteq \{D_1, \ldots, D_m\}$.

$Definition\ 2 : (K\text{-}Anonymity\ Property)$ Table T is k-anonymous with respect to attributes $D_1, \ldots, D_m$ if every unique tuple $(d_1, \ldots, d_m)$ in the (multiset) projection of T on $D_1, \ldots, D_m$ occurs at least k times. That is, the size of each equivalence class in T with respect to $D_1, \ldots, D_m$ is at least k.

3.2 Metrics for information loss

The information loss has a wide concept and various metrics have been proposed in privacy preserving data analysis. In order to maintain the utility of soccer fitness data, we should change the table as small as possible. That is, the information loss after anonymity should be minimized.

The first metric we use is one that attempts to capture in a straightforward way the desire to maintain discernibility between tuples as much as is allowed by a given setting of k. The $discernibility\ metric$ [5] can be mathematically stated as follows:

$$C_{DM}\left(T^{'}\right) = \sum_{1<i<m} \mid E \mid^2 \qquad (1)$$

In this expression, the set E_i refers to the equivalence class of tuples in table $T^{'}$ induced by the anonymization. The number of the equivalence classes is m.

Another interesting cost metric we use was originally proposed by Xu [8]. On a numeric attribute D_i, the normalized certainty penalty is defined as $NCP_{A_i}(t) = \frac{z_i - y_i}{|A_i|}$, where the $\mid A_i \mid = max - min$ is the range of all tuples on attributes A_i . Then the $normalized\ certainty\ penalty\ metric$ can be formally defined as follows:

$$C_{NCP}\left(T^{'}\right) = \sum_{1\leq i\leq m} NCP\left(t_i\right) \qquad (2)$$

where the t_i is the i-th record, and the m represent the number of attributes.

4 Anonymity method

In this section, we consider methods to protect the soccer fitness data publishing from linking attack. In this paper, two algorithms for achieving k-anonymity are applied in the soccer fitness data.

4.1 The greedy partitioning algorithm

LeFevre $el.\ al.$ [7] transform the k-anonymity problem into a partitioning problem. The approach consists of two phases. At the first step, multidimensional regions are defined that cover the domain space by finding a partitioning of the d-dimensional space, where d is the number of quasi-identifier attributes, so that each partition contains at least k tuples. And in

the second step, the records in each partition are generalized such that they all share the same quasi-identifier value. The solution of strict partitioning is Algorithm 1.

Algorithm 1 Top-down greedy algorithm for strict multidimensional partitioning

Input: A table T and an integer k.
Output: An anonymized table T'.
1:Anonymize($partition$).
2:if(no allowable multidimensional cut for $partition$)
3: return ϕ: partition $\rightarrow$ generalization.
4: else
5:$dim \rightarrow choose\_dimension()$;
6:$fs \leftarrow frequency\_set(partition, dim)$;
7:$splitVal \leftarrow find\_median(fs)$;
8: $lhs \leftarrow \{t \in partition : t.dim \leq splitVal\}$;
9:$rhs \leftarrow \{t \in partition : t.dim > splitVal\}$;
10:return Anonymize(lhs) $\bigcup$ Anonymize(rhs).

4.2 The K-Means clustering algorithm

We can also transform the k-anonymity problem into a clustering problem. The clustering problem is to find a set of clusters from a given set of n records such that each cluster contains at least k $(k \leq n)$ data points and that the sum of all intra-class distances is minimized and the inter-class is maximized. Using the k-means clustering algorithm, it is generated in the following three steps.

Step 1: Clustering()

In this step, we use the Algorithm 2 to cluster the soccer fitness data into classes. In the Algorithm 2, the line 1 and 2 is the initialization process, we calculate the number of the classes and randomly select m records as the center of each class. At the line 3 to line 8, there is the iteration process. At each iteration, we will calculate the distances between each tuple t and each class c, then we will pick out the minimal distance and add the tag of the class to it. Finally, the clustering centers will be recalculated. During this process, euclidean distance is used to measure the similarity of each record and clustering centers.

Algorithm 2 Clustering Algorithm

Input: A table T an integer k.
Output: The clustered table T'.
1:$m \leftarrow \mid T \mid /k$;
2:$center\_list \leftarrow random\_select(m)$;
3: while $!is\_center\_stable\,(center\_list)$ do
4: foreach $t \in T$ do
5: foreach $c \in center\_list$ do
6: $j \leftarrow find\_min\_distance(t, c)$;
7: $add(t, c_j)$;
8: $center\_list \leftarrow reCalcul(center\_list)$;
9: return the clustered table T'.

Step 2: Merging()

After the clustering, there may exists some groups whose records are less than k. In order to solve this issue, we introduce a merging processing. At the process, the similarity of two class centers will be calculated by the euclidean distance, and then merge the small

group into the cluster, whose distance between them is minimal such that the records' number will be larger than k.

Step 3: Anonymizing()

Finally, we anonymize the records in the same class so that have the same quasi-identifier value. The procedures of the last step is Algorithm 4.

Algorithm 3 Anonymizing Algorithm
Input: The merged table T''.
Output: An anonymized table T'''.
1: foreach $cluster \in T''$ do
2: foreach $quasi - attr \in cluster$ do
4: $sum \leftarrow CalculAll(quasi - attr)$;
5: Replace it with the $mean$;
6: return the anonymized table T'''.

5 Experiments

In this section we evaluate the performance of the proposed k-anonymity algorithms. The experiments are conducted on a 2.93 GHz Intel(R) Core(TM) 2 Duo CPU with 4GB running the Windows 7 operating system.

5.1 Dataset

We use the real soccer fitness data, which is collected form the soccer players in the ShanDong LuNeng from 2016 to 2017 five quarters. All the tuples in the table have 12 attributes. Among them, the athlete's name is replaced by a pseudonym, one nominal attribute and numeric attributes are contained. In this paper, we just use the height and weight as the quasi-identifiers, and the others are sensitive attributes, which can reveal some potential privacy information about the players.

5.2 Information loss

DM cost vs k(Fig. **2***)* Fig. **2** shows the relative changes of the DM cost with three methods by varying k . The reason for this is that as k increases, more and more records are generalized to have the same quasi-identifier value, resulting in greater loss of information. In addition, the clustering algorithm is better than the Mondrian. This is because that the number records in each partition is larger than $2k$ when the k is small. And the cost of the clustering is gradually close to the Mondrian when the k increasing.

*NCP cost vs k (*Fig. **3***)* Fig. **3** further gives the experimental result of the relation between NCP and k . NCP cost is mainly used to measure the degree of generalization of records, which can be adopted as a good metric of data utility. In this paper, we calculate the percentage of the NCP. The NCP generally increase as k increases, so it exhibits some trade-off between data privacy and data utility. The Mondrian(strict) is better than the clustering and Mondrian(relax), and sometimes Mondrian(relax) exists the same percentage of ncp because of the same partitions when the k is close.

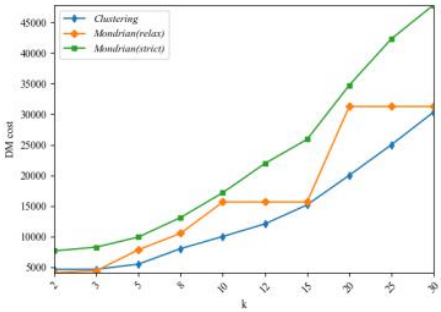

Fig. 2. The relation between $DM\ Cost$ and k.

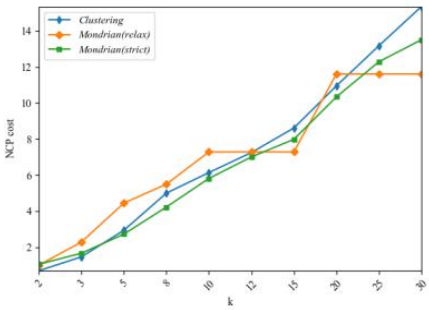

Fig. 3. The relation between $NCP\ Cost$ and k.

6 Conclusions

In this paper, we apply the partitioning-based and the clustering-based algorithms in the real-world privacy soccer fitness data publication to achieve the k-anonymity model. Simultaneously, we use the discernibility metric and normalized certainty penalty metric as effective metrics for information loss. Experimental results show that these methods can preserve the soccer players' privacy and guarantee the utility of these data. Yet, we also think that there are quite a few promising works to be done in the further research. First, the current data was published one time, but if we want to publish another collected data, it will be an incremental data release problem.

References

1. Benjamin C M Fung, Ke Wang, Rui Chen, and Philip S Yu. Privacy preserving data publishing: A survey of recent developments. ACM Computing Surveys, 42(4):14, 2010.
2. Alessio Rossi, Luca Pappalardo, Paolo Cintia, Marcello Iaia, Javier Fernandez, and Daniel Medina. Effective injury prediction in professional soccer with gps data and machine learning. 2017.
3. Latanyasweeney. k-anonymity: A model for protecting privacy. International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 10(05):557–570, 2002.
4. LATANYA SWEENEY. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05):571–588, 2002.
5. Roberto J Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In International Conference on Data Engineering, 2005. ICDE 2005. Proceedings, pages 217–228, 2005.
6. Kristen Lefevre, David J. Dewitt, and Raghu Ramakrishnan. Incognito: efficient full-domain k-anonymity. In ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, Usa, June, pages 49–60, 2005.
7. Kristen Lefevre, David J. Dewitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In International Conference on Data Engineering, pages 25–25, 2006.
8. Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Wai Chee Fu. Utility-based anonymization using local recoding. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–790, 2006.